Ethics and Autonomous Systems: Perils and Promises

Ronald C. Arkin

Artificial intelligence (AI) and its role in autonomous systems have promised everything from utopian freedom to existential dystopia. The unfilled hyperbole surrounding past and present promises regarding AI futures has left many people skeptical, afraid or just confused. Rational discussion is often left in the wake due to the fears and fantasy evoked by the press and Hollywood. Fortunately, as a byproduct, this has resulted in a blossoming of worldwide discourse on the ethical implications of the intelligent machines we are creating. Many near- and mid-term ethical concerns have arisen with the advent of autonomous systems: particularly regarding driverless cars, privacy and drones, companion and intimate robotics, the displacement of jobs by intelligent machines, and warfighting robots among others. The IEEE Global Initiative on the Ethics of Autonomous Systems, the United Nations, the International Committee of the Red Cross, the White House, and the Future of Life Institute are among many responsible organizations that are now considering the ramifications of the real world consequences of machine autonomy as we continue to stumble about trying to find a way forward.

One thing can clearly be stated: We are creating autonomous technology faster than we are able to (1) understand its implications; (2) interpret it within moral frameworks; and (3) create policy and legislation to govern its development and deployment. Progress on AI, despite a rather slow pace for decades, finally appears to be accelerating as evidenced by advances in machine learning (Google's AlphaGo), cognitive computing (IBM's Watson), robotics (Boston Dynamics' MiniSpot and Atlas), speech understanding (Apple's Siri, Amazon Echo)… the list goes on. While we are now in a catch-up phase regarding regulation and legislation, society and governments need to be far more proactive and must discuss and debate the difficult questions surrounding the use of artificial intelligence. If we ignore the increasingly rapid pace of advances, we do so at our own peril as the very fabric of our society and international relations will be tested at the very least and possibly ruptured in unpredictable ways at the worst.

There are generally no universal rights or wrongs with respect to autonomous systems given that there are competing ethical frameworks by which to assess their outcomes. This is further compounded by cultural and societal differences worldwide. Tensions exist between the rights of individuals or groups (embodied in rights-based/Kantian ethical theories) versus maximizing the overall happiness of all concerned (as found in consequentialist/utilitarian theories). Nonetheless, policy and law must follow as a result of such deliberations.

I am not concerned about the posited existential threats to humanity from artificial intelligence and the associated apocalypse [Time 2014]: the sky is not falling. We will have more than ample opportunity to destroy ourselves by other

means prior to the singularity[1] should it ever occur. While I am glad smart people are thinking about it, the present holds far more perils to humanity in my mind than this futuristic hypothetical fear.

In this light, let's review three of these near-term critical threats from the point of view of a practicing roboticist and ethicist of late who's been involved in these discussions internationally for over a decade.

Driverless cars – Who lives and who dies?

In many ways this is the topic du jour given the expected proliferation of self-driving cars in the near future. It has even been stated by some that children born today will not ever drive. The motivation is clear – humans are the most dangerous things on the road, and replacing them with autonomous AI could lead to a saving of life and reduction of injuries. Driverless cars are immune to DUI, distracted driving, and road rage – many of the issues leading to highway accidents. They can offer the elderly, the blind, and otherwise physically challenged mobility where now they have none. The National Highway Traffic Safety Administration is issuing guidelines (not law however) in the summer of 2016 for autonomous cars[2].

The core ethical questions for driverless cars are twofold[3]. The first is the classic case of the trolley problem that is fodder for almost all basic ethics classes (i.e., who lives and who dies when a choice must be made in an unavoidable accident). The most straightforward example is when an autonomous vehicle recognizes that a crash is inevitable for whatever reason – what should it be programmed to do? Expose the driver to the maximum risk to protect others in the vicinity? Veer out of the way and possibly take the lives of other car occupants or pedestrians while protecting the driver? Who makes this decision for the car? Where does liability rest?

Second, should the automobile always obey the law to the letter? This has already resulted in the Google car being rear-ended when it came to a legal full stop at a stop sign[4]. It can also result in potential road rage and dangerous driving by people irritated by a vehicle following the speed limit exactly. When accidents and even deaths occur, such as the recent Tesla fatality when the car was in autopilot mode,

---

[1] Roughly speaking, the posited point when machine intelligence exceeds human intelligence.

[2] http://www.digitaltrends.com/cars/nhtsa-autonomous-vehicle-guidelines/ accessed 7/15/2015.

[3] These questions are discussed in more detail in a recent IEEE Spectrum article: http://spectrum.ieee.org/transportation/self-driving/can-you-program-ethics-into-a-selfdriving-car accessed 7/15/2015.

[4] http://www.bloomberg.com/news/articles/2015-12-18/humans-are-slamming-into-driverless-cars-and-exposing-a-key-flaw accessed 7/15/2015.

liability will ultimately end up being defined in the courtroom, and subsequently law will be either enforced according to current standards or changed as a result.

Using robots to save noncombatant lives: A Moral Imperative?

The arguments surrounding the use of lethal force by autonomous systems in the battlefield has been raging for over a decade. Should robots be empowered to take human life? Could they yield a reduction in collateral damage and save civilian lives? Does their use violate human dignity at some level?[5] The United Nations in Geneva has been debating this for over 3 years, and it is unclear whether a total ban, a set of regulations, or anything will result. Part of the problem is definitional: what is a lethal autonomous robot? Some such as myself, argue that they exist and have for decades, while others say they have not been created yet. By my definition autonomous systems are simply the next generation of precision-guided munitions. Others argue that these systems will ultimately bear responsibility, resorting to the philosophers' definition with respect to their possessing free will or moral agency – which is certainly not the case now (or perhaps ever).

Almost everyone supports the notion of meaningful human control, but we cannot all agree on exactly what that means. The good news is that these discussions are ongoing at the highest levels of international discourse and may result in changes to International Humanitarian Law in terms of regulations or even a ban if deemed necessary. But progress towards a consensus is slow at best and may never emerge. To me the fundamental problem is with respect to non-combatant casualties: the status quo is utterly unacceptable and technology can, must, and should be used to address this problem. Parallels can be drawn with respect to the driverless car argument: here human beings are the most dangerous things in the battlespace with respect to civilians. Warfighters are on occasion prone to poor judgment, carelessness, or even atrocities in their use of force. Something must be done to better protect noncombatant life. And if it's not battlefield robots that don't experience fear, anger, and frustration as humans do; that can process more information from more sources faster; that can assume far more risk on behalf of noncombatants than any human soldier in their right mind would; that may be able to adhere better to International Law and the Rules of Engagement better than human warfighters; then we must find some other way to assure greater safety for civilians. Simply doing nothing continues to leave the innocent in grave peril.

Intimate robotics – how close is too close?

Robot sex and intimacy is well outside of the bounds of most civil discussion, unfortunately. Why this is unfortunate is that the technology is already emerging without any real discourse. There is a need for basic scientific multidisciplinary research accompanied by open and frank discussion, as currently these machines

[5] ACM held a plenary debate on this topic in 2015, which resulted in two opposing position papers in CACM [Goose 2015, Arkin 2015].

are being developed and deployed in an ethical vacuum. A colleague once pointed to the fact that DVRs were propelled by their serving as a vehicle for pornography, that the internet gained widespread usage due to pornography, and that robotics technology is the next stage of this revolution of sexual mores. Sexual toys and artifacts are not new - they have been present with humanity since ancient times. But with intimate robotics, we are referring to systems that actively foster attachment and, yes, even love in the user that is directed towards the intelligent artifact. Humans have a natural propensity for developing these artificial relationships, already evidenced by fondness towards cars, the often excessive caring of technological objects such as AIBO Sony's robot dog, the Tamagotchi handheld digital pet, and even Roombas [6]. Sherry Turkle writes eloquently about this in her book [Turkle 2014], but stops before intimacy enters the picture. Levy's landmark book *Sex and Love with Robots* captures the state of the art about a decade ago, with a significant focus on cultural differences with respect to acceptance.

There are few forums to even discuss these issues – a conference workshop in Malaysia was declared illegal in 2015[7]. Governmental funding for studying these issues is a complete nonstarter. Yet we as a society are fascinated with the possibility. Science fiction has written on this for years … Asimov's Robots of Dawn, among others (1983). It is also evidenced by recent films such as *Her* and *Ex Machina*, and others such as *Metropolis, AI, Blade Runner, Cherry 2000*, and TV series such as *HUM∀NS* and *Battlestar Galactica*.

Reiterating, real technology of this sort is beginning to be created without meaningful ethical discourse. As intimate robotics as a field progresses without discussion, more and more of society will come under its influence. What is the effect on human-human relationships as this technology progressively becomes more mainstream? [Borenstein and Arkin 2015]. If we do not attend to this, the results again may stray far beyond any of our expectations.

What now?

I have tried not to be too prescriptive in my discussion of these issues, as it is not the place for a roboticist to tell the world what is right and wrong. It is my place, however, to state that these are ethical quandaries that need to be discussed now. I have reviewed but a few of the issues confronting us today with autonomous systems moving into the real world. We need not be fearful, but we need to be proactive in understanding the societal impact of this technology before policy generation and legislation. We are already well engaged in proactive discussions regarding lethal autonomous systems. For driverless cars the ethical discussion is

[6] http://www.popsci.com/technology/article/2010-03/emotional-attachment-roombas-suggests-humans-can-love-their-bots-seriously accessed 7/15/2016.
[7] http://www.bbc.co.uk/newsbeat/article/34615532/love-and-sex-with-robots-conference-cancelled-in-malaysia accessed 7/15/2015.

concurrent with the introduction of the technology into the marketplace with uncertain results regarding liability and responsibility. With respect to intimate robotics we are not really engaged at all in the necessary ethical discussions that will guide our acceptance or rejection of the technology.

It is up to all of us to secure a reasonable future for ourselves, our families, our society and the world. Technologists need to engage in these discussions and be circumspect on the technology they are creating. Proactive discussion is essential – start today.

References

Arkin, R.C., "The Case for Banning Killer Robots: Counterpoint", *Communications of the ACM*, Vol. 58, No. 12, pp. 46-47, 2015.

Asimov, I., *Robots of Dawn*, Doubleday, 1983.

Borenstein, J. and Arkin, R.C., "Robots, Ethics, and Intimacy: The Need for Scientific Research", *2016 Conference of the International Association for Computing and Philosophy (IACAP 2016)*, Ferrara, IT, June 2016.

Goose, S., "The Case for Banning Killer Robots: Point", *Communications of the ACM*, Vol. 58, No. 12, pp. 43-45, 2015.

Luckerson, V., "5 Very Smart People Who Think Artificial Intelligence Could Bring the Apocalypse", Time, Dec. 2, 2014. http://time.com/3614349/artificial-intelligence-singularity-stephen-hawking-elon-musk/

Turkle, S., *Alone Together: Why We Expect More from Technology and Less from Each Other*, Basic Books, 2014.